Formulating Simple Structured Queries using Temporal and Distributional Cues in Patents

Le Zhao and Jamie Callan
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh PA 15213

Abstract

Patent prior art retrieval aims to find related publications, especially patents, which may invalidate the patent. The task exhibits its own characteristic because of the possible use of a whole patent as a query. This work focuses on the use of date fields and content fields of the query patent to formulate effective structured queries. Retrieval is performed on the collection of patents which also share the same structure as the query patent, mainly priority dates, application date, publication date and content fields. Unsurprisingly, results show that filtering using date information improves retrieval significantly. However, results also show that a careful choice of the date filter is important, given the multiple date fields existent in a patent. The actual ranking query is constructed based on word distributions of title, claims and content fields of the query patent. The overall MAP of this citation finding task is still in the lower 0.1 range. An error analysis focusing on the lower performing topics finds that the citation finding task (given publication recommend citations, which is a very similar setup as this year's prior art evaluation) can be very different from the prior art task (finding patents that invalidates the query patent). It raises the concern that just the citations included in query patents can be a biased and incomplete set of relevance judgements for the prior art task.

1 Introduction

TREC 2009 Chemical track [1] consists of two main tasks, Technical Survey and Prior Art search. Both tasks retrieve documents from a collection of mostly chemical patents. This paper reports our participation in the Prior Art search task.

In the real world, the prior art search scenario starts with a query patent, usually a full patent with priority dates, claims, full content of the patent and citations. The search system used by a patent officer should identify previous patents or publications which may invalidate the query patent.

Unlike a typical ad hoc retrieval scenario, there are two components of the prior art search request, first is *prior* and second is art, *relevant* work. We used date filtering in Indri query language [3] to address the first aspect, and weighted query terms extracted from title, claims and the body text of the query patent to address the second aspect.

In this year's Chemical track, the Prior Art task is a bit different from the real world setting.

First, the citation list of the patent is assumed hidden from the system (in order to be used as automatic relevance judgements). This absence of citations makes it impossible to study how patents cite prior art and how to make use of the citations to improve retrieval.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2009		**00-00-2009 to 00-00-2009**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Formulating Simple Structured Queries using Temporal and Distributional Cues in Patents	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Carnegie Mellon University,Language Technologies Institute,School of Computer Science,Pittsburgh,PA,15213**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009) held in Gaithersburg, Maryland, November 17-20, 2009. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**8**	

Second, possibly a difference both from real world prior art search and from ad hoc retrieval, the query as a patent is itself contained in the collection to retrieve matches from. This means the retrieval engine has to remove the query itself from the top matches in order to get reasonable evaluation results.

Third, no human judgements were used, only entries from the citation list of the query patent and other sources (which include around 20-25 cited patents) are treated as relevant results, and everything else as irrelevant. This is a sparse set of judgements compared to other more typical TREC tasks, and possibly also a biased set with the authors of the query patent and a few patent officers as the only contributors to the list. The fact that patent officers are expert searchers in the domain, does not mean that the list is complete. It would be worthwhile to investigate on a few topics to verify the quality of this judgement set. With many un-judged documents in the result list, it is very hard to make conclusions such as whether some retrieval method is effective, simply because the top ranked un-judged results could be irrelevant as assumed by the evaluation, but could also be relevant. The evaluation only gives a lower bound of the true performance.

2 Query Formulation

2.1 Date Filtering: Which date matters

A patent contains multiple date entries. From earlier to later, there may be and typically are multiple priority dates each corresponding to some earlier patent owned by the same assignees. The use of the priority dates is to extend the priority of the patent claims to some earlier time, so that only patents earlier than the priority date could invalidate the current patent. There is also an application date and a publication date of a patent (if it passes the prior art check of the patent office).

With all these dates, the only ones that matter are the priority dates (if no priority dates then the application date) of the query patent and the publication dates of the prior art patents.

For the multiple priority dates, we compare two strategies.

- F_early: Filter out documents published after earliest priority date.
- F_late: Filter out documents published after latest priority date.

Of the two filtering strategies, F_early is more restrictive. However, it is possible that a patent published after the earliest priority date of the query patent can still invalidate some claims of the query patent. The reason is that these invalidated claims are not extended by the priority patents to the earliest priority date in the list of priority dates.

F_late is the safest, yet tightest criterion in terms of simple date filtering, i.e. without considering the actual content of the claims. To construct a more restrictive date filter than F_late and at the same time without missing relevant patents, an automatic approach will have to take into account both the date of the priority claims and their content, so that the system could find out which claims are extended to which priority date. Patent officers do this for some high profile cases.

An extra advantage of date filtering is that it automatically filters out the query patent itself, since the publication date will be later than the application and priority dates.

2.2 Content distribution for ranking

After date filtering, relevance ranking yields the final ranked list of prior art patents. For relevance ranking, any content, such as title, claims, abstract and the descriptions of the query patent can be used. We base our retrieval query on the title of the query patent, and consider two variations,

- R_titleclaim: Weighted combination of the title and terms from the claims.
- R_titleclaimdesc: Weighted combination of the title and terms from the claims, with the terms from the claims weighted by their occurrences in the other text fields of the query patent.

Relevance ranking strategy R_titleclaim only takes the title and claims of the query patent to construct the retrieval query, and strategy R_titleclaimdesc also takes into account the other text fields of the query patent such as the abstract and descriptions.

In theory, for R_titleclaimdesc, we would want to use all the terms from the query patent's content. However, that gives too many terms in a single query. This kind of huge queries presents a challenge for the typical search engine, both time-wise and space-wise. Here, we chose to only include the terms in the title and claims, but weighted them according to their respective occurrences in the content fields of the query patent. This strategy contains the time and memory consumption, and also gives a close approximation of the ideal case.

2.3 Example query

Take the following query patent for example (the XML schema is simplified for this presentation):

<patent>
<title>example patent</title>
<claim id=1>example claim one</claim>
<claim id=2>example claim two</claim>
<abstract>example abstract</abstract>
<description>example description</description>
<publication-date>2003/05/20</publication-date>
<application-date>2002/01/20</application-date>
<priority-date>2002/01/20</priority-date>
<priority-date>2001/01/20</priority-date>
</patent>

In Indri query language, two example queries are shown below, corresponding to the two strategies R_titleclaim and R_titleclaimdesc.

- **Q1**: #weight(w_1 #combine(example patent) w_2 #weight(2 example 2 claim 1 one 1 two))
- **Q2**: #weight(w_1 #combine(example patent) w_2 #weight(4 example 2 claim 1 one 1 two))

Here, #weight is the weighted combination operator, with w_1 and w_2 being the weights for the two components being combined. #combine is a uniformly weighted combine operator. We assume $w_1 + w_2 = 1$.

Applying filtering strategies F_early and F_late respectively to Q1 gives

- **Q3**: #filrej(#dateafter(2001/01/20) **Q1**)
- **Q4**: #filrej(#dateafter(2002/01/20) **Q1**)

Here, #filrej is the filter rejection operator, which rejects any document that matches the filtering criterion. In Q3 all documents that have a date field after 2001/01/20 will be filtered out.

Table 1. Evaluation of best performing parameters on training sets

Dataset	Ranking query	Filter strategy	Weight on title	MAP	MRR	P@5
EP	R_titleclaim	No filter	0	0.0115	0.0926	0.0533
		F_early	0	0.0175	0.0941	0.0533
		F_late	0	0.0442	**0.2393**	**0.1200**
	R_titleclaimdesc	F_early	0	0.0245	0.1405	0.0400
		F_late	0	**0.0481**	0.2377	0.1067
US	R_titleclaim	No filter	0.6	0.0586	0.1483	0.0933
		F_early	0.6	0.0863	0.2862	0.1067
		F_late	0.6	0.1083	0.3239	0.1200
	R_titleclaimdesc	F_early	0.2	0.1186	**0.3459**	**0.2133**
		F_late	0.2	**0.1309**	0.3452	**0.2133**

Table 2. Official evaluation of the submitted runs on both full 1000 topics and top 100 topics

Runs	Ranking & Filtering	title Weight	Test Set	MAP	MRR	P@5	NDCG	bpref	R@100
CMU09Chmtcd	R_titleclaim + F_late	0.6	PA1000	0.0647	0.4004	0.2140	0.2344	0.3605	0.1678
			PA100	0.0517	0.3002	0.1280	0.2090	0.3779	0.1631
Unofficial	R_titleclaim + F_late	0	PA1000	0.0715	0.4334	0.2414	0.2517	0.3756	0.1754
			PA100	0.0222	0.1340	0.0660	0.1086	0.2007	0.0780
CMU09Chmtcdd	R_titleclaimdesc + F_late	0.2	PA1000	**0.0975**	0.5129	0.2994	**0.3091**	0.4570	**0.2326**
			PA100	**0.0894**	**0.3819**	**0.2100**	**0.2972**	0.4999	**0.2610**
Unofficial	R_titleclaimdesc + F_late	0	PA1000	0.0961	**0.5130**	**0.3004**	0.3055	0.4510	0.2262
			PA100	0.0883	0.3703	0.2080	0.2894	0.4813	0.2583
Unofficial	R_titleclaimdesc + F_early	0.2	PA1000	0.0784	0.4534	0.2458	0.2565	0.3889	0.2000
			PA100	0.0794	0.3731	0.1820	0.2736	0.4636	0.2474
Unofficial	R_titleclaimdesc + no Filter	0.2	PA1000	0.0765	0.2901	0.1910	0.3069	**0.5115**	0.2079
			PA100	0.0620	0.2097	0.1260	0.2752	**0.5562**	0.2285
Runs from All groups	Mean of each query		PA1000	0.0279	N/A	N/A	0.1639	0.3614	0.0594
			PA100	0.0229	N/A	N/A	0.1525	0.3950	0.0654
Runs from All groups	Max of each query		PA1000	0.1835	N/A	N/A	0.4192	0.6602	0.3375
			PA100	0.1688	N/A	N/A	0.4359	0.7432	0.4055

3 Experiments

We used Indri search engine of the Lemur toolkit [3] to index all the patent documents for the Prior Art task. The index includes all date fields in the patent documents and allows the retrieval application to query them using Indri query language.

Two datasets were used for tuning the parameter w_1 in our models. First dataset is the TREC official 15 EP patent topics with their citations, and a second unofficial set of 15 US patents prepared by the University of Iowa group [2]. Because test patents are mostly US patents, the parameter was trained only on the US topic set. Because of the small number of training topics that can be used, we intentionally kept our methods simple and robust so that it would generalize to larger datasets with ease.

We present evaluations of the best performing parameters on the two train sets, EP and US, respectively. For testing, two data sets were used, one being a set of 1000 patent topics, and another being the first 100 topics of the 1000 topic set, we call them PA1000 and PA100. Besides the listed parameter, smoothing parameter μ for Dirichlet smoothing is tuned on the EP set and fixed at 10000. Results of the tuned models are shown in Table 1. As shown in Table 1, The US set is more responsive to parameter tuning. Because the test set is consisted mainly of US patents, we used the parameter values trained on the 15 US topics.

Test performance is shown in Table 2. As expected, the performance on the test collection correlated better with the US training set than the EP set. The parameters tuned on the US set do perform better than being tuned on the EP set, when tested on the PA set.

Filtering improves performance about or over 100% compared to no filtering on the training set and about 20% on test set. The best performing filter is to filter out publications after the latest priority date. Maybe because of the timeliness of patents, filtering by the earliest priority date yields a significant decrease in performance. The interval between the two dates is typically in the range of 0 to 3 years. Thus, the actual interval for a query patent would affect how the date filter performs.

The typical ranking query contains a title of around 10 words and words from claims which vary from 50 to over 300 unique terms. Using the description of the query patent does not increase the length of the query, but only affects the weight of the terms. A relatively short example is shown below, with both ranking and filtering parts and using only the title and claim fields of the query patent.

```
#filrej( #dateafter(07/07/1994)
        #weight( 0.6 #combine( detergent compositions)
                0.4 #weight(
                        16 1 14 bleaching 12 agent 11 composition 11 oxygen 11 7 10 4 8 u 8 2 7 o 7 available 6
                        claims 5 triazacyclononane 5 silver 5 coating 5 clo 5 organic 5 3 4 mn 4 minutes 3 co 3
                        mniii 3 0 3 5 3 bispyridylamine 3 description 3 n 3 containing 3 described 3 releasing 3
                        method 2 mixtures 2 time 2 compound 2 mixture 2 dentate 2 remainder 2 rate 2 mniv 2
                        source 2 tri 2 making 2 sprayed 2 intimate 2 completely 2 oac 2 cl 2 trimethyl 2 selected 2
                        premixed 2 bleach 2 dispersing 2 compositions 2 pf 2 released 2 perchlorate 2 oil 2 10 2 di
                        2 group 2 methyl 2 release 2 non 2 cobalt 2 consisting 2 interval 2 process 2 paraffin 2
                        particles 2 present 1 claim 1 perhydrate 1 nh 1 salt 1 copper 1 total 1 corrosion 1 bispyridyl
                        1 chlorate 1 bi 1 8 1 dry 1 measured 1 partially 1 mnivbipy 1 och 1 trisdipyridylamine 1
                        comprises 1 mnivn 1 isothiocyanato 1 ligands 1 combination 1 triglycerides 1 bis 1 amine 1
                        6 1 bipy 1 binuclear 1 pyridylamine 1 mniiimniv 1 relasing 1 inorganic 1 mixed 1 precursor
                        1 iron 1 hydrogenated 1 peroxyacid 1 additional 1 inhibitor 1 tetra 1 tris 1 level 1
                        derivatives 1 provided 1 diglycerides 1 gluconate 1 mono 1 wholly 1 complexed 1
                        catalyst ) ) )
```

On both training and test sets, using the description to reweight terms from title and claims not only improves retrieval performance a lot, but also stabilizes performance on the test set. This is a simple and efficient way of re-formulating the structured query.

No matter what ranking strategy is in use, date filtering always improves retrieval performance by a similar amount. This means, the filtering strategy and the ranking strategy are quite independent in improving retrieval performance. If we think of time as one facet of the query patent, other facets might bring similar independent retrieval performance improvements.

With filtering and ranking combined, the retrieval performance is still in the lower range of 0.1 in MAP and only 0.2-0.3 in P@5. Even if we take the maximum among all runs, the MAP is still below 0.2. These figures are lower than a typical ad hoc retrieval run. It could be because the task is difficult; with the especially long queries, there is a greater chance of finding false positives. Another possible reason is that there are in fact more relevant documents than those included in the evaluation, and those were treated simply as irrelevant, causing MAP etc. measures to be lower than it would be. Thus, the measures here presented are lower bound estimates of the true measures. More manual judgements need to be done to see how complete the current relevant document lists are, and how tight the lower bounds are.

4 Conclusions

Date filtering improves retrieval performance a lot. The best filtering strategy requires result patents to be published earlier than the latest priority date of the query patent. This strategy defeats other filtering strategies by a large margin. Further improvement over the automatic filtering approach would require consideration of both claim content and priority-claim dates. Date filtering also improves retrieval performance independent of the ranking strategy.

Using the content description of the query patent together with the title and claims improves retrieval performance significantly. Using description also stabilizes performance on test sets. The large set of 1000 test topics now available would allow more techniques to be tested on formulating more effective ranking and filtering queries given the query patent.

5 Error Analysis

Because the overall performance is still low, less than 0.1 in MAP, an error analysis was performed to investigate.

First, in Figure 1, we show retrieval performance of individual topics on the two training sets. As we can see, only a few topics get a reasonably high average precision, while a lot of the other topics have a performance close to 0. For two of those near 0 topics in the EP set, we looked at the top ranked false positives and top ranked relevant documents.

EP topic 3 (patent number EP-0690122) is about an *Oxygen-releasing (controlled release) bleaching agent*, with a non-paraffin oil organic silver coating agent, and additional corrosion inhibitor compound[1]. Top 6 ranked documents are non-relevant, they are US-5114611, US-5314635, US-5194416, EP-0544490, US-5227084 and EP-0530870, and they are about *controlled release laundry bleach products*, or *bleach activation*. Following these

[1] This summary is an excerpt from the “Brief Description of the Invention” or “Summary of Invention” in other patents. We find this field to be informative when starting to read the patents, and understanding the patents.

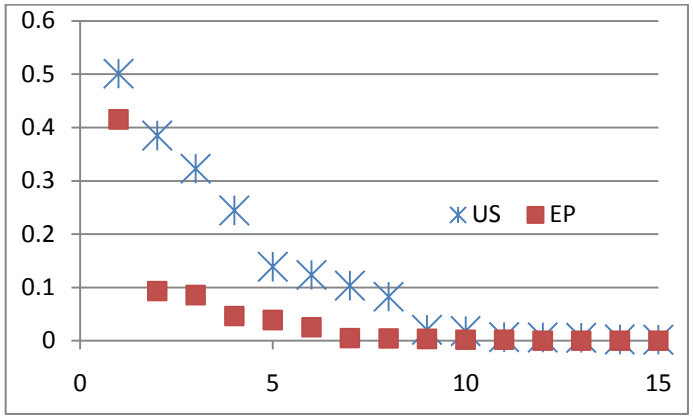

Figure 1. Average precision of individual topics on the two training sets.

false positives are two relevant results, EP-0458398 and EP-0549272, both about *bleach activation* in laundry use. From non-chemist's eyes, the relevant and false positives are not any different in terms of relevance.

Why are the relevant ones relevant? Because these two relevant documents come from the citations of the query patent, a more specific question to ask is, in what context are these two patents cited, and why cited? First, both citations were created by the applicant, as annotated in the <patcit> fields of the query patent, meaning they are not the ones that a patent office found, that would invalidate the patent. Second, looking at the content of the query patent, the two citations were made as examples of bleach activation catalysts which is one small part of how the bleaching agent works. Quoting from the query patent: "Preferred examples of these catalysts include … and mixtures thereof. Others are described in European patent application publication no. 549,272." Again, for a non-expert on bleaching agents, if these two patents are considered relevant, then the false positives can be easily included and justified in the query patent as citations. We further looked at the other citations made by the applicants and conclude that the actual citation list prepared by the patent applicants can be a very biased and incomplete set of prior art. This might be a consequence of the common citation practice of patent applicants, e.g. citing just enough patents to make the case clear instead of citing every relevant patent, and focusing on the differences with the cited patents instead of commonalities.

For EP topic 2, the situation is very similar, a biased set of marginally relevant patents were cited by the patent applicants.

This means, the Prior Art task of this year's Chemistry track is more like a citation (related work) finding task. And citation finding is very different from a real world prior art search situation where the goal is to find all patents that would possibly invalidate the query patent. In terms of evaluation design, a more realistic setup would be to have some level of manual judgements on the top ranked un-cited patents, so that we have a more accurate idea of what the real false positive rates are. Currently the task treated only cited patents as relevant, which we show to be potentially biased and incomplete. A double check of the cited patents might also be needed to ensure the quality of the relevance judgements.

6 Acknowledgements

We thank patent expert Aleksandr Belinskiy for clarifying the automatic date filtering strategy, and Christopher Harris and colleagues from University of Iowa for creating and sharing the US patent 15 topic test set.

This work is supported by National Science Foundation grant IIS-0707801 and IIS-0534345. The views and conclusions are the authors', and do not reflect those of the sponsor.

7 References

[1] TREC 2009 Chemical track. https://wiki.ir-facility.org/index.php/TREC_Chemistry_Track. Retrieved Oct 20, 2009

[2] Yelena Mejovac, Viet Ha Thucc, Steven Fosterd, Christopher Harrisa, Bob Arensc, Padmini Srinivasan. TREC Blog and TREC Chem: A View from the Corn Fields. TREC 2009 Notebook.

[3] INDRI - Language modeling meets inference networks. http://www.lemurproject.org/indri/. Retrieved Oct 1, 2009